

**School of Computing and Information Systems**

**ISSS616 Applied Statistical Analysis with R**

**Project Report**

**Statistical Insight and Analysis on Los Angeles Crimes Rates 2020 - 2021**

**Group 6:**

Antonius Handy

Jonathan Yang Zixian

Ng Wee Kien

Phan Tuan Minh

Widya Tantiya Yutika

**Table of Contents**

[1. Introduction 2](#_Toc99049414)

[2. Overall Concept 2](#_Toc99049415)

[3. Data 3](#_Toc99049416)

[4. Descriptive Statistics 4](#_Toc99049417)

[4.1 Shiny App Design Considerations 4](#_Toc99049418)

[4.2 Crime Occurrence 4](#_Toc99049419)

[4.3 Victim Demographic Distribution 6](#_Toc99049420)

[4.4 Geospatial Mapping 7](#_Toc99049421)

[4.5 Crime Over Time 9](#_Toc99049422)

[4.6 Top Crimes and Top Police Station Areas 9](#_Toc99049423)

[5. Inferential Statistics 9](#_Toc99049424)

[5.1 Shiny App Design Considerations 9](#_Toc99049425)

[5.2 Hypothesis Test 10](#_Toc99049426)

[5.3 Confidence Interval Test 11](#_Toc99049427)

[5.4 Analysis of Variance (ANOVA) Test 11](#_Toc99049428)

[5.5 Chi-Squared Test of Association 12](#_Toc99049429)

[6. Conclusion 13](#_Toc99049430)

[6.1 Limitations 13](#_Toc99049431)

[6.2 Assumptions 13](#_Toc99049432)

[6.3 Recommendations 13](#_Toc99049433)

[6.4 Future Work 14](#_Toc99049434)

# Introduction

Crime is one of the most concerning societal issues in every nation, generating substantial costs to society. According to numerous prior studies, there are four fundamental costs caused by crimes such as victim costs, criminal justice system costs, crime career costs, and intangible costs (i.e. psychological distress, fear, pain, and suffering). These intangible costs are equally concerning to society, if not paid enough attention, may threaten the social and democracy stability of a country and cause a domino effect that increases the number and severity of crimes.

In Los Angeles, the Los Angeles Police Department (LAPD) plays an important role in safeguarding the lives and property of the people, reducing the incidence and fear of crime, and enhancing public safety while working with the diverse communities to improve their quality of life[[1]](#footnote-2). While LAPD aims to achieve a city free of crime and public disorder as closely as possible, the Los Angeles City Attorney creates crime prevention strategies, measures, and policies to improve the quality of life which in turn can reduce the number of crimes.

*“In Los Angeles, the City Attorney plays a leading role in shaping the future of Los Angeles city by fighting to improve the quality of life in LA’ neighbourhoods, reducing gang activity, preventing gun violence, standing up for consumers and elderly, protecting the environment and so much more[[2]](#footnote-3).”*

-Los Angeles City Attorney Office

During a Los Angeles Police Department (LAPD) event in January 2022, Mayor Eric Gargetti and Police Chief Michael Moore revealed that the number of crimes in 2021 had increased as compared to previous years, especially in homicides (up 143%) and shootings (up 12%), reaching its highest point in the last 15 years[[3]](#footnote-4). Another report made by Fox 11 also claimed that the number of crimes increased in 2021[[4]](#footnote-5).

The increasing trend points toward the opportunities to improve crime prevention strategies and measures to reduce the risk of crimes occurring and their potential harmful effects on individuals and society.

# Overall Concept

Prior to creating crime prevention strategies, it is essential for both LAPD and the City Attorney to have a fundamental understanding of the crime patterns based on past records, which can be used for future predictions.

The objective of this project is to perform exploratory data analysis on crime that has occurred in LA over the past two years and subsequently provide insights on the crime patterns. This is aligned with one of the City Attorney Mike Feuer’s newest plans to identify the patterns, trends and community factors that lead to gun violence[[5]](#footnote-6). Furthermore, this project will also confirm the validity of the news by performing some hypothesis tests.

Text

Description automatically generated with low confidence*Figure 1: Landing Page of the Shiny Crime LA App*

The interactive Shiny dashboard will provide a data visualisation experience that can enable the City Attorney to create effective policies which will target the corresponding demographic groups and communities.

# Data

The data set used is provided by the Los Angeles Police Department and extracted via the Los Angeles’ Open Data website[[6]](#footnote-7). The dataset reflects incidents of crime in Los Angeles, transcribed from original crime reports, from the period 2020-2021.

The dataset comprises 403,598 unique crime incidents, with up to 28 data fields for each incident. Using the existing dataset, we also derived 8 additional data fields for ease of usage. Some further processing was done upon importing the data into RStudio, where we binned the ages of victims into nine groupings.

data **<-** read\_csv**(**"crime\_data\_v1.csv"**)**

# Keeping only 2020 and 2021 data

data **<-** filter**(**data, Year %in% c**(**2020, 2021**))**

#Victim Age grouping

age\_group\_label **<-**c**(**"1-9","10-19","20-29","30-39","40-49","50-59","60-69","70-79",">=80"**)**

age\_group **<-** cut**(**data**$**Vict\_Age, breaks**=**c**(**0,9,19,29,39,49,59,69,79,**Inf)**, labels**=**c**(**"1-9","10-19","20-29","30-39","40-49","50-59","60-69","70-79",">=80"**))**

#Adding new column Victim Age Group into the data

data**$**age\_group **<-** age\_group

Code Snippet 1: Reading and Processing Data in RStudio

The key data fields used for our study can be categorised and summarised as follows:

|  |  |
| --- | --- |
| **Data Field** | **Description** |
| DR\_NO | Unique case number for each case incident |
| DATE\_OCC, TIME\_OCC  Day, Month, Year, Time, Hour (derived) | Date and time of crime incident |
| AREA | Area code and name of police station responding to the crime incident |
| Part 1-2  Crm\_Cd\_Category | Classification of crime incident |
| Vict Age, Sex, Descent | Descriptions of the victim by age, gender and ethnicity |
| Premis Cd  Premis Desc | Code and description of the premise of the crime |
| Weapon Used Cd  Weapon Desc | Type of weapon used in the crime |
| LAT  LON | Longitude and latitude of location of crime |

*Table 1: Data Field and Description*

# Descriptive Statistics

## Shiny App Design Considerations

Users are greeted with 3 collapsible tabs under the ‘Exploratory Analysis’ page. The design objective of this section is to allow users to explore the data, look for interesting patterns, and start forming ideas and/ or hypotheses which can be further tested under the ‘Inferential Analysis’ section.

Background pattern

Description automatically generated

*Figure 2: Layout of Exploratory Analysis Tab*

Users are first encouraged to get a sensing of the overall dataset under the *‘Basic Exploratory Analysis Charts – all data’* tab. Equipped with some background context of the available data, they are then able to dive deeper into different sub-populations, by modifying filters under the *‘Reactive Inputs’* tab. These reactive filters work in conjunction with the last tab *‘Deep Dive Exploratory Analysis Charts - selected categories’*, where interactive charts are updated in real time, to empower users in generating specific ideas which may be noted for further tests.

## Crime Occurrence

To visualise timings for when crimes typically occur, we created a heat map showing the day-of-week and time-of-day for which crime occurs. Users are also able to filter the heat map for specific crime types using our Shiny dashboard.

The heat map results (Figure 1) provided us with some interesting insights on timings of crime occurrences. 12 noon is the peak period for crime incidents, and this is pronounced on Mondays, Wednesdays and Fridays. Furthermore, Fridays are the most popular day for crimes to be committed. Majority of crime incidents happens from 12 noon onwards and slow down after midnight, where most people are off the streets and in their beds.

When filtering by crime types (Figure 2), the results showed that each crime have different peak periods – for example theft is more likely to happen during the day, burglary and theft from vehicle is more likely to happen at night, while rate of assault case is spread across the whole day.

Using the results, crime prevention strategies with a focus on days and timing where crime occurrences are more prevalent can be deployed more effective in reducing overall crime rates. The results can also be used to allocate police manpower resources effectively. By filtering down into crime type, the strategies can also focus on preventing certain major crimes.

A picture containing chart

Description automatically generated

Figure 3: Heat Map of Crime occurrences by Day of Week and Time of Day

Calendar

Description automatically generated with medium confidence

Figure 4: Heat Map of Each Crime Occurrences by Day of Week and Time of Day

## Victim Demographic Distribution

Understanding the victim demographics (age, gender, and race) can help identify vulnerable communities that may require increased police protection. Using the *ggplot* function, we were able to create bar graphs showing the distribution of victims by age group, gender and race. The insights we gained from this analysis were also used for further inferential statistics testing, which is elaborated later in our report.

**Age Group –** Most victims of crime in Los Angeles fall within the 20-29 and 30-39 age groups, this result is not surprising as it matches the LA census data from census reporter website[[7]](#footnote-8).

Chart, bar chart

Description automatically generated Chart, histogram

Description automatically generated

Figure 5: Victim Age Group Distribution Figure 6: LA Population Age Group Distribution7

**Gender** – There were more crimes committed against male than female or undisclosed gender. There are some victimless crimes which is excluded from this analysis.

Chart, bar chart

Description automatically generatedChart, pie chart

Description automatically generated

*Figure 7: Victim Age Group Distribution Figure 8: LA Population Age Group Distribution7*

**Race** – Crimes against Hispanics/Latinos/Mexicans were the highest, and this is in line with the population of LA, with Hispanics being largest ethnic group (48.2%) in the city. Whereas crimes against Vietnamese and Black people was the second and third highest respectively, even though each group made up less than 12% (Vietnamese is a part of Asian) and 8% of total population in LA. This could be evidence of a community being more vulnerable to crime and warranting increased attention for police protection by the LAPD.

Chart

Description automatically generated

Figure 9: Victim Descent/Race Distribution

Chart, waterfall chart

Description automatically generated

Figure 10: LA Population Descent/Race Distribution7

## Geospatial Mapping

Using *leaflet* and the LON and LAT data fields, our Shiny dashboard can display crime incidents at the location where crime was committed. Users can use this interactive mapping tool to zoom into different areas, filter by date (month, year) and also select to focus on certain crime types (eg. burglary, rape etc.).

This serves as a tool for LAPD to identify crime hot spots within the city, while city planners can consider this information when making re-zoning plans for the city.

Overall\_Vict\_Descent**<-**data %>%

df**<-**data%>%

filter**(**Lat**!=**0**)**%>%

filter**(**Crm\_Cd\_Category**==**"Rape"**)** %>%

filter**(**Area\_Name**==**"Central"**)**

df**$**popup**<-**paste**(**"<b>Report Number #: </b>", df**$**Dr\_No, "<br>", "<b>Category: </b>", df**$**Crm\_Cd\_Category,

"<br>", "<b>Description: </b>", str\_to\_sentence**(**df**$**Crm\_Cd\_Desc, locale **=** "en"**)**,

"<br>", "<b>Day of week: </b>", df**$**Day,

"<br>", "<b>Date: </b>", df**$**Date\_Occ,

"<br>", "<b>Time: </b>", df**$**Time,

"<br>", "<b>Police Station Area Name: </b>", df**$**Area\_Name,

"<br>", "<b>Longitude: </b>", df**$**Lon,

"<br>", "<b>Latitude: </b>", df**$**Lat**)**

leaflet**(**df, width **=** "100%"**)** %>% addTiles**()** %>%

addTiles**(**group **=** "OSM (default)"**)** %>%

addProviderTiles**(**provider **=** "Esri.WorldImagery",group **=** "World Imagery"**)** %>%

addMarkers**(**lng **=** **~**Lon, lat **=** **~**Lat, popup **=** df**$**popup, clusterOptions **=** markerClusterOptions**())** %>%

addLayersControl**(**

baseGroups **=** c**(**"OSM (default)", "World Imagery"**)**,

options **=** layersControlOptions**(**collapsed **=** **FALSE)**

**)**

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated*Code Snippet 2: Leaflet Package for Geospatial Mapping of Crime Occurrences*

Map

Description automatically generated

Figure 11: Interactive Mapping Tool for Crimes in LA

## Crime Over Time

The blue line chart displays the evolution of number of crimes over time. Using the *geom\_smooth* function*,* the smoothed conditional mean and regression line is shown for ease of seeing patterns/trends.

A picture containing text, device, antenna, different

Description automatically generated

*Figure 12: Number of Crimes Over Time*

## Top Crimes and Top Police Station Areas

This section provides user with information on the top crime types and the police station areas which handle the crimes. The slider bar allows the users to view the number of bar graphs according to their needs. However, if the total crime types/police station areas after filtering is less than 5, it will only show the filtered result as a default and so the slider bar can be ignored.

Graphical user interface, application

Description automatically generated

*Figure 13: Top Crime Types and Top Police Station Areas*

# Inferential Statistics

## Shiny App Design Considerations

Similar to the ‘Exploratory Analysis’ tab, the design is kept clean and simple for users. Users are able to first adjust the reactive inputs/ filters to their liking, where it will be used for the subsequent sections covering Confidence Intervals, Hypothesis Testing, ANOVA, and Chi-Squared Test of Association.

Background pattern

Description automatically generated

*Figure 14: Layout of Inferential Analysis Tab*

## Hypothesis Test

Referring to the news made by Fox 11, hypothesis test for two independent population is performed to test its validity. This will enable us to understand whether there is a difference in daily crime rates between 2020 and 2021.

Fifty random samples are taken from the population and sigma values are known. Assuming a significance level α = 5%, the null hypothesis and alternative hypothesis are as follows.

**Null hypothesis H0:**

**There is no significant difference in daily crime rates between 2020 & 2021**

**Alternative hypothesis H1:**

Text

Description automatically generated**There is a significant difference in daily crime rates between 2020 & 2021**

*Figure 15: Two-tail Test Result*

Two-tail test result shows that p-value is higher than 5%, therefore we do not have sufficient evidence to claim that there is a significant difference in daily crime rates between 2020 and 2021.

Graphical user interface

Description automatically generated with medium confidence

*Figure 16: Hypothesis Test Result in Shiny*

Similarly, our Shiny App design methodology allows users to update the reactive inputs, which are used to filter the data fed into the hypothesis tests. Furthermore, users are also able to tweak test parameters such as the Confidence Level, Sample Size, and Comparison Direction (‘Less’, ‘Greater’, ‘two.sided’). These changes are reflected in real time. The output of the Z-test has been extracted and simplified to include key statistics while making the conclusion straightforward to users. There may be many claims and/ or assumptions made about crime, and this design allows users to test ideas/ claims formed from earlier sections in a flexible and efficient manner.

## Confidence Interval Test

Text

Description automatically generatedFollowing the result of hypothesis testing, we also constructed confidence intervals on the daily average crime incidents in 2020 vs 2021. Similar to hypothesis testing, 50 random samples are taken from the population. At 95% confidence level, the confidence interval for daily average crime rates in 2020 was between 516 to 556, while this range increased to 542 – 568 in 2021. The narrowness of the confidence interval indicates a greater degree of precision and can be used accurately by LAPD to estimate the number of crime incidents that will be committed daily in LA. The results can then be used to optimise the LAPD’s police manpower allocation strategies.

*Figure 17: Confidence Interval for* *Daily Crime Rates*

Chart

Description automatically generated with low confidence

*Figure 18: Confidence Interval Results in Shiny Crime LA*

In addition to the selection of data under the *‘Reactive Inputs – Main’* tab, users are also able to tweak the confidence level and sample size parameters and observe the updated Lower and Upper bounds of the confidence interval in real time.

## Analysis of Variance (ANOVA) Test

ANOVA test is used to compare multiple means. We are interested in comparing the average of monthly crime rates across different genders, namely male (M), female (F) and undisclosed (X).

Ten random samples are taken from the population. At significance level of 5%, the null hypothesis and alternative hypothesis are as follows.

**Null hypothesis H0:**

**The averages of monthly crime rates across different genders are equal**

**Alternative hypothesis H1:**

**The averages of monthly crime rates across different genders are not equal**

*Text

Description automatically generatedFigure 19: ANOVA and Tukey’s Test Result*

It is observed that the p-value is less than the significance level 0.05, therefore we reject the null hypothesis and conclude that there is no sufficient evidence to state that the averages of monthly crime rates across different genders are equal. There is at least one gender having a different average from the other genders.

Graphical user interface, application, table

Description automatically generatedTukey’s test is then performed to find out which gender is different from each other. When M is compared to F, the p-value is less than 0.05 which means their averages are significantly different. The average difference is known to be 2229.5 which indicates that males are more vulnerable to be the victim of crimes compared to females. Similar analyses are also done when comparing X and F as well as X and M. In conclusion, the monthly crime rates are at the highest level for males, followed by females and lastly undisclosed.

*Figure 20: ANOVA Test Results in Shiny Crime LA*

Like previous sections, the filtered dataset (from the reactive inputs earlier) are put through an ANOVA test. Users are able to tweak the sample size (aggregated at the monthly level) and observe the results of the various pairs within the chosen demographic variable. 4 of the more prominent variables (Gender, Crime Code, Race, and Area Name) have been catered for.

## Chi-Squared Test of Association

Further work was done within Shiny Crime LA to allow users to observe for any potentially significant associations between the key demographic variables. The team has chosen to provide for comparisons between Crime Code Category against 3 other demographic variables (Race, Area Name, and Gender). This allows users to make certain inferences both using the entire dataset, but also to deep-dive into granular sub-categories.

At 5% significance level, the null hypothesis and alternative hypothesis are:

**Null hypothesis H0:**

**No association exists between crime category and demographic variable**

**Alternative hypothesis H1:**

Graphical user interface, text, application, email

Description automatically generated**Association exists between crime category and demographic variable**

*Figure 21: Results from Chi-Squared Test of Association*

The app extracts the key results, and users are told the significance of the generated p-value from the Chi-Squared test of association. If the p-value is less than 5%, we reject the null hypothesis and say that there is an association between those two variables. Otherwise, we conclude that there is no association found.

If association exists, users can easily read the cross-tab table and see which pairs of variables where crime tends to lean towards, or away from.

# Conclusion

## Limitations

There are two major limitations in this project: scope and geography.

Scope: Our data available is only for 2 years (2020 to 2021), and so it is difficult to observe the trend before COVID-19 period. Given time and available sources, we would like to expand the data to earlier years to have a full and clear view of crime rates trend and patterns in the city.

Geography: Data available is only from the Los Angeles Police Department, and therefore the observation, insights, and analysis are valid only for Los Angeles area. While LA is a major city and an important metropolitan, it is difficult to assume that the insights we draw from the data can apply to other cities across the US.

## Assumptions

Our assumption on the data is that all crimes happened independently of each other, and no crimes are paired.

## Recommendations

Through our study, we were able to identify key recommendations that the LAPD and City Attorney can use to formulate city’s crime prevention strategies:

* Our tests found that crime affects different victim demographics disproportionately, especially the Latino/Hispanic community, which has significantly higher rate of being a victim of crime compared with other races. LAPD can look at this data and investigate the root cause to identify vulnerable people and areas that may require increased police protection.
* There are identifiable crime patterns in terms of location (eg. Downtown LA) and timing (eg. 12 noon on Wednesdays) – LAPD can use this information to plan their manpower allocation as well as potential relocation of police stations for faster response time.

## Future Work

Future work could focus on:

* Ways to update the LA crime dataset in an automated manner through the usage of any available APIs, that can feed directly into a real-time interactive dashboard
* To weave in secondary sources of data relevant to the demographics in LA, and further explore any interesting insights.
* To expand the scope to a larger area such as the entirety of the US, where conclusions can be drawn on the US population instead of LA-specific.

1. <https://www.lapdonline.org/mission-statement/> [↑](#footnote-ref-2)
2. <https://www.lacityattorney.org/about> [↑](#footnote-ref-3)
3. <https://spectrumnews1.com/ca/la-west/public-safety/2022/01/13/homicides--vehicle-thefts--officer-involved-shootings-increased-in-la-in-2021> [↑](#footnote-ref-4)
4. <https://www.foxla.com/news/stats-show-crime-is-spiking-in-los-angeles> [↑](#footnote-ref-5)
5. <https://www.lacityattorney.org/post/la-city-attorney-mike-feuer-proposes-eight-point-plan-to-tackle-gun-crimes-in-city> [↑](#footnote-ref-6)
6. <https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8> [↑](#footnote-ref-7)
7. <https://censusreporter.org/profiles/16000US0644000-los-angeles-ca/> [↑](#footnote-ref-8)